Accepted Article

Received Date: 25-Sep-2015

Revised Date: 17-Jan-2016

Accepted Date: 28-Jan-2016

Article Type: Article

A framework for inferring biological communities from environmental DNA

Andrew Olaf Shelton[1*]

James Lawrence O'Donnell[2]

Jameal F. Samhouri[1]

Natalie Lowell[2]

Gregory D. Williams[3]

Ryan P. Kelly [2]

[1] Conservation Biology Division, Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic & Atmospheric Administration, Seattle, WA 98112

[2] University of Washington, School of Marine and Environmental Affairs, 3707 Brooklyn Ave NE, Seattle, WA 98105

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1890/15-1733.1
This article is protected by copyright. All rights reserved.

[3] Pacific States Marine Fisheries Commission, Under contract to Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic & Atmospheric Administration, Seattle, WA 98112

* Corresponding author: ole.shelton@noaa.gov

Keywords: environmental DNA, community surveys, Bayesian statistics, ecosystem assessment, multinomial-Poisson transformation, quantitative PCR.

Abstract

Environmental DNA (eDNA)—genetic material recovered from an environmental medium such as soil, water, or feces—reflects the membership of the ecological community present in the sampled environment. As such, eDNA is a potentially rich source of data for basic ecology, conservation, and management, because it offers the prospect of quantitatively reconstructing whole ecological communities from easily-obtained samples. However, like all sampling methods, eDNA sequencing is subject to methodological limitations that can generate biased descriptions of ecological communities. Here, we demonstrate parallels between eDNA sampling and traditional sampling techniques, and use these parallels to offer a statistical structure for framing the challenges faced by eDNA and for illuminating the gaps in our current knowledge. Although the current state of knowledge on some of these steps precludes a full estimate of biomass for each taxon in a sampled eDNA community, we provide a map that illustrates potential methods for bridging these gaps. Additionally, we use an original dataset to estimate the relative abundances of taxon-specific template DNA prior to PCR, given the abundance of

Accepted Article

This article is protected by copyright. All rights reserved.

DNA sequences recovered post-PCR-and-sequencing, a critical step in the chain of eDNA inference. While we focus on the use of eDNA samples to determine the relative abundance of taxa within a community, our approach also applies to single-taxon applications (including applications using qPCR), studies of diversity, and studies focused on occurrence. By grounding inferences about eDNA community composition in a rigorous statistical framework, and by making these inferences explicit, we hope to improve the inferential potential for the emerging field of community-level eDNA analysis.

Introduction:

A central aim of ecology is to understand the distribution and abundance of organisms, which requires estimates of the occurrence, density, or biomass of the organisms in natural populations. Whether counting individuals in a habitat, in a population, or across an assemblage, making inferences about an entire community from an observed subset of individuals is fundamental to ecological science. Unfortunately all sampling techniques are potentially subject to bias, undermining accuracy and confidence in estimates of critical ecological parameters. Visual surveys may overlook or misidentify cryptic species, surveys that capture individuals with nets or traps may under-represent small or elusive prey, and quadrat-sampling methods for non-mobile flora and fauna can underestimate the abundance of rare species or miss landscape-scale patterns. Fortunately there is a large and sophisticated literature dedicated to examining and improving efficacy and reducing bias for a range of sampling problems for terrestrial, marine, and aquatic systems (Cochran 1977, Royle and Nichols 2003, Cotter and Pilling 2007, Elith and Leathwick 2009). In this paper, we contribute to this literature by developing a general statistical

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

framework as well as specific statistical sampling methods for the emerging field of environmental DNA.

Recent advances in molecular biotechnology have resulted in the emergence of a new survey tool, whereby the DNA present in an environmental medium (such as soil or water; hereafter environmental DNA or eDNA), can be used to infer the presence of organisms nearby (Jerde et al. 2011, Yoccoz 2012). There are currently two distinct molecular approaches for eDNA. In the first, the amount of a known DNA sequence - presumably from a single taxon - is determined from quantitative polymerase chain reaction (qPCR; Thomsen et al. 2012, Nathan et al. 2014). The second approach is to amplify some suitable region from all genomes present in a sample using PCR, and sequence the resulting products (amplicons) using massively parallel sequencing technologies, without *a priori* knowledge of the organisms present or their genetic sequences (e.g. Ventner et al. 2004, de Barba et al 2015, Leray and Knowlton 2015). While the qPCR approach is being used in several applications to monitor rare or invasive species (Lodge et al. 2012; Turner et al. 2014), such methods can involve extensive development for each taxon of interest, and cannot easily provide insight into community-level patterns. Sequencing methods could feasibly provide relative abundance data for a suite of species in the community, as the relative proportions of taxon-specific DNA sequences observed may reflect the relative proportions of DNA in the environment (Yoccoz 2012). While attempts have been made to link sequence counts to biomass (e.g. Evans et al. 2015), no such study has yet evaluated the complex chain of processes and associated uncertainty linking these two states (Iversen et al. *In Press*). Thus, one barrier to the widespread adoption of the sequencing approach is the lack of formal methods for linking this new data type (counts of DNA sequences) to the underlying pattern of interest (the abundance or biomass of taxa comprising a community; Yoccoz 2012).

This article is protected by copyright. All rights reserved.

Accepted Article

Conceptually, using eDNA to infer the biomass or abundance in a community is largely analogous to traditional non-molecular methods. Figure 1 illustrates how eDNA and traditional sampling attempt to provide information about the same quantity: the biomass of each species in the environment. Both eDNA and traditional sampling aim to make inferences about distinct stages that are potentially measurable (latent states; represented by boxes in Fig. 1), and processes which transform one stage to the next (arrows in Fig. 1).

Before turning to sampling methods for eDNA, we first describe a general theoretical framework in terms of traditional sampling of a marine fish community, with the goal of quantifying the biomass of each taxon. Common sampling methods for fish communities include using a variety of net technologies (trawl, gillnets, cast nets, seines, etc.), systems using baited hooks, and visual surveys. Importantly, the process of inference from data using any of these methods can be conceptualized using the diagram in Figure 1. We use fish communities as an example with which we are familiar, and for which there is a long history of explicitly modeling uncertainty, but the larger point applies to all ecological sampling.

For example, for a sample collected using a trawl net, the total biomass of fish taxon *i* at a particular location, *l*, and a particular time, *t*, $B_{i,l,t}$ is a function of the biomass or counts observed in the net, $F_{i,l,t}$ (Fig. 1). Given that we only observe $F_{i,l,t}$ the process of estimating $B_{i,l,t}$ from $F_{i,l,t}$ can be written as a conditional quantity, $[B_{i,l,t}|F_{i,l,t}]$. For expositional purposes, we simplify notation by assuming a single sample time and location, $[B_i|F_i]$. As Fig. 1 shows, the biomass in the environment (B_i) is not connected to the biomass captured by the net (F_i) by a single process but rather a chain of distinct processes. A full description of the sampling process would explicitly include each step. For example, researchers commonly extract a subsample of

This article is protected by copyright. All rights reserved.

individuals (E_i) from the full catch of the net (D_i) to determine the taxon-specific count (F_i), which itself may be influenced by taxonomic identification errors or other processes (Fig. 1).

From this conceptual framing it should be clear that our estimate of the taxon's biomass B_i is influenced by at least three sets of processes: (1) the sampling approach to obtain the collection D_i, (2) the methods used to reduce the full collection to the subsample E_i, and (3) the identification and enumeration methods that result in the taxon specific count F_i. Statistically, we can expand the inference of interest $[B_i|F_i]$ into three conditionally independent processes (for general discussion of conditional modeling see Clark 2007, Cressie and Wikle 2011)

$$[B_i|F_i] = [B_i|D_i][D_i|E_i][E_i|F_i]\ . \qquad (1)$$

Thus, any estimate from sampling data must implicitly or explicitly make assumptions or estimate these three components. For example, the second term on the right side, $[D_i|E_i]$ describes the proportion of the total catch taken in a subsample. If the entire catch is included, $[D_i|E_i]$ = 1, and this term can be dropped from the model.

While accounting for $[D_i|E_i]$ is relatively straightforward, other terms in eq. 1 are more difficult. Indeed, determining how biomass present in the environment corresponds to the total catch in the net, $[B_i|D_i]$, is a classic and persistently difficult problem that has been explored extensively in ecology (Royle and Nichols 2003, Elith and Leathwick 2009) and fisheries (see the fisheries concepts of "catchability", and "selectivity"; Beverton and Holt 1957 section 8, Arreguín-Sánchez 1996, Venebles and Dichmont 2004). For our hypothetical marine fish example, the mesh size, design, and deployment of the net, among other characteristics, will interact with the true density of each species to determine which are captured (the quantity $[B_i|D_i]$ in eq. 1; Beverton and Holt 1957 section 8, Arreguín-Sánchez 1996). Similar challenges face the determination of $[E_i|F_i]$; individual skill and experience will affect the efficacy and

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

accuracy of taxonomic identification. Our purpose here is merely to note that such complexities plague virtually all sampling problems—whether terrestrial or marine, from the poles to the equator.

The basic inferential framework introduced above (eq. 1, Fig. 1) readily applies to the problem of reconstructing ecological communities from eDNA. Below, we outline the processes connecting ecological communities to observations of eDNA, and briefly summarize the state of knowledge about each process. We then construct a statistical model for analyzing community eDNA data that accounts for some of the processes that can potentially bias inference from eDNA data and provide a worked example for applying these methods to a marine eDNA dataset. We end by briefly discussing further methodological needs for eDNA data and making recommendations for best practices. Throughout, we focus on the use of eDNA for community sampling and highlight the inferential and empirical connections between traditional and eDNA sampling methods.

Conceptual models for eDNA

Here we derive a model structure to estimate the relative amount of biomass present in a community for some set of taxa of interest, by sampling eDNA. While we develop the framework in the context of estimating abundance for multiple species from sequenced DNA, both models of occurrence (e.g. Ji et al. 2014) and of single species abundance (e.g. Jerde et al. 2011) are special cases in our framework, as will be discussed later. Our general approach also applies to qPCR methodologies. We focus on the detection and quantification of taxa that are not directly sampled. For example, if we collected a liter of water from the environment, we focus primarily on inferring the abundance of fishes, invertebrates, and mammals from individual cells

This article is protected by copyright. All rights reserved.

Accepted Article

(and accompanying DNA) contained in that water sample. While similar methods could be applied to bacteria and other microorganisms that can be directly measured and sequenced from a small sample, we do not specifically address such cases here; direct sequencing rather than PCR based approaches may be more appropriate for small, abundant taxa (Yu et al. 2012).

To derive a general model for eDNA we need to explicitly consider the data in hand and the process that led to the observation of the data. We assume that a researcher has collected a sample of seawater—although soil, fecal, or other samples are essentially equivalent—for the purposes of recovering eDNA from an ecological community. After filtering the sample, extracting total DNA, and amplifying the DNA of interest using oligonucleotide PCR primers, we observe counts of unique DNA sequences from a high-throughput sequencer (e.g., Illumina, 454, Ion Torrent). Note that there are many possible molecular methods by which the data can be derived. For all cases, though, the number of observed DNA sequences for each type is a function of: 1) the true, but unknown, density of DNA of each taxa present in the water, 2) the amount of DNA captured on the filter and subsequent DNA extraction, 3) the primer set and its interaction with the DNA sequence of each taxon present, 4) the number of PCR cycles performed, 5) the error rate of the sequence analyzer, and myriad other factors. In short, the observed counts of DNA sequences are a complicated stochastic realization of the true amount of DNA present in the environment for each taxa. While eDNA protocols can be designed to minimize such stochastic forces, they cannot be eliminated altogether. Defensible ecological inference therefore depends upon identifying and estimating the parameters that may substantially influence observed counts of DNA sequences.

By analogy with the net sampling example, the process by which biomass is translated into DNA sequences matched to taxonomic groups is probabilistic (Fig. 1). Specifically, the

This article is protected by copyright. All rights reserved.

biomass of each taxon must be translated through several intermediate states before it is observed as counts of DNA sequences. For taxon *i*, let W_i be the density of DNA in the environment, X_i be the amount of DNA collected from the environment, Y_i be the DNA present after DNA extraction, and Z_i be the DNA sequences recovered. We acknowledge that there are other reasonable ways of parsing the process of generating and making inference from eDNA (i.e. the framework we discuss here is extendable, and additional states could be added to Fig. 1). However the latent states in Fig. 1 are intuitive and, potentially, directly measurable with existing technologies.

As in eq. 1, the amount of biomass B_i estimated from eDNA sampling is the product of four conditionally independent steps,[1]

$$[B_i|Z_i] = [B_i|W_i][W_i|X_i][X_i|Y_i][Y_i|Z_i] \qquad (2)$$

Information about each link in this inferential chain is required to properly infer B_i from observed counts of DNA sequences that emerge from a DNA sequencer Z_i. Such information can be some combination of prior information about the processes connecting these latent states, direct observations of the states, and biologically justified assumptions about each component. There are two corollaries of this point: *i*) any inferences made about B_i from eDNA make implicit and/or explicit assumptions about the other components on the right side of eqn. 2; and, *ii*) if there is no information about any of the components on the right side of eq. 2 (or researchers are unwilling or unable to make assumptions about these components), it will be impossible to make inference about B_i from eDNA observations alone. A parallel problem arises frequently in fisheries; biologists are unwilling to assert that the actual biomass is mirrored by

[1] For the remainder of the manuscript, we let capital Roman letters denote random variables, lowercase roman letters denote realizations of random variables, and Greek letters denote parameters. Bold lowercase denote vectors and bold uppercase are matrices.

This article is protected by copyright. All rights reserved.

Accepted Article

observed catches (i.e. the connection between *B* and *D* in Fig. 1 cannot be bridged). Therefore survey catches are frequently used as indices of abundance not estimates of absolute abundance (Kimura and Zenger 1997, Cotter and Pilling 2007). Despite not reflecting actual abundance, such indices play a critical role in fisheries, wildlife sciences, and management (Branch et al. 2010, Jannot and Holland 2013). The formulation of eq. 2 also serves to point out where information is missing and to motivate future research on poorly understood topics (Yoccoz 2012, Pedersen et al. 2015).

Other structures for Figure 1 are reasonable and we encourage investigators to modify the chain of inference represented in Figure 1 to meet their specific sampling needs. We view Figure 1 not as a rigid form for analyzing eDNA but as a framework which can be modified to suit individual purposes and clarify thinking about the inferences that can and cannot be drawn from available eDNA data. We expect improved and more complex analytical structures to be developed for eDNA as the technology and its use evolve.

An important point of Figure 1 is that the traditional sampling and eDNA arms of the figure are only connected through the true biomass, *B*, represented at the top of the figure. This structure serves to remind investigators that that directly comparing eDNA and traditional sampling data is fraught with difficulty and can only be logically done with a full sampling model for both how counts of OTUs observed from a sequencer (*Z*) connect to biomass (*B*) and how traditional sampling observations connect to biomass. Alternatively one could make strong assumptions about the connection between *Z* and *B.* Indeed the most difficult step for both eDNA and traditional sampling in marine environments is the first step in each pathway (between *B* and the density of DNA in the environment *W*, and between *B* and collected individuals in a traditional sample, *D*; Fig. 1). To date, we know of no eDNA study which has explicitly linked *B*

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

and W under field conditions and very few that have linked them under controlled laboratory conditions (e.g. Takahara et al. 2012, Thomsen et al. 2012). To date, most researchers have either asserted that the proportion of sequences observed from environmental samples mirror the abundance (either count or biomass) of physically collected individuals or, alternatively, concluded proportions of sequences are proportional to abundance based on visual inspection (for example, see de Vargas et al. 2015, their Figs. W2B and W2C and accompanying text). While these correlations may accurately reflect a functional link between individuals in the environment and eDNA, we would point out that a complex and diverse set of processes that separate D and Y mean that there are large number of ways to arrive at spurious correlations between these two states. It is therefore desirable to explicitly assess each link in the inferential chain linking observed DNA sequences to biomass or some other biological/ecological parameter of interest.

While eq. 2 is instructive to broadly frame eDNA problems, the processes that connect the latent states must be detailed to make this model useful in practice. Specifically, the rates of transition between the states presented in Fig. 1 are controlled by parameters that do not appear in eq. 2; we introduce those parameters here. Let $\boldsymbol{\theta_i}$ be a set of species-specific parameters associated with transition from B_i to W_i (e.g. DNA shedding (Klymus et al. 2015, Iversen et al. *In Press*) and degradation (Thomsen et al. 2012, Strickler et al. 2015; Fig. 1), $\boldsymbol{\phi_i}$ be taxon-specific parameters associated with transition from W_i to X_i (e.g. the small scale patchiness of DNA in the water), $\boldsymbol{\psi_i}$ be taxon-specific parameters associated with DNA filtering and extraction (the transition from X_i to Y_i), and $\boldsymbol{\xi_i}$ define taxon-specific parameters associated with PCR amplification and sequencing driving the transition from Y_i to Z_i (e.g. the match of a primer

This article is protected by copyright. All rights reserved.

sequence to the DNA input to template DNA sequence). Eq. 2 can be rewritten to include these parameters for all taxa simultaneously,

$$[\boldsymbol{B}|\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\xi}] = [\boldsymbol{B}|\boldsymbol{W},\boldsymbol{\theta}][\boldsymbol{W}|\boldsymbol{X},\boldsymbol{\phi}][\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\psi}][\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{\xi}] \qquad (3)$$

To connect these equations to empirical observations, they must be matched to appropriate likelihood functions; we demonstrate in detail how to do so in the section "***Statistical methods for community eDNA***" below.

It bears noting that the current state of knowledge with respect to eDNA limits our ability to estimate all terms on the right-hand side of eq. 3, although at least some data are available from which to begin such estimation. Here we briefly summarize the state of knowledge with respect to each term in eq. 3 (Fig. 1).

1) *Processes in the transition from biomass, B, to DNA present in the environment, W (*$\boldsymbol{\theta}$*)*

- DNA shedding rates are positively correlated with biomass and influenced by diet (Takahara et al. 2012, Kelly et al. 2014, Klymus et al. 2015, Evans et al. 2015) and ambient eDNA density varies by species (Thomsen et al. 2012). Small DNA fragments (ca. 100 base pairs) degrade within a few days in the marine environment (Thomsen et al. 2012) but in some cases DNA signals are detectable for weeks to months (Barnes et al. 2014, Strickler et al. 2015). DNA shedding and degradation rates likely differ among taxa and among life-stages (Maruyama et al. 2014, Iversen et al. *In Press*), though these differences are not well studied.
- In aquatic environments transported DNA does not appear to accumulate downstream from the organism shedding it (Laramie et al. 2015) but rather remains at similar concentrations downstream over short distances (Pilloid et al. 2014). DNA may be moved over longer distances by bulk flow (Deiner and Altermatt 2014) or by mobile predators that transport prey DNA in their gut and deposit it in their feces (Merkes et al. 2014).

This article is protected by copyright. All rights reserved.

2) Processes in the transition from DNA in the environment, W, to DNA collected on a filter, X $(\boldsymbol{\phi})$*, and from DNA collected on a filter, X, to DNA present after extraction, Y* $(\boldsymbol{\psi})$

- Although methods for capturing eDNA influence the amount of useful sequence data, they likely do not cause taxon-specific biases (Feinstein et al. 2009, Turner et al. 2014, Deiner et al. 2015). However, pre- and post-processing sample storage and DNA extraction methods can produce taxon-specific biases (Carrigg et al. 2007, Deiner et al. 2015).

3) Processes included in the transition from DNA present after extraction, Y, to DNA present after sequencing, Z ($\boldsymbol{\xi}$*)*

- PCR amplification of multi-taxon DNA samples introduces sequence-specific biases due to differential primer binding strength (Lee et al. 2012); to a lesser degree the number of PCR cycles may exacerbate these biases (Polz and Cavanaugh 1998, Sipos et al. 2007).
- To improve cost efficiency by increasing sample throughput, a unique nucleotide sequence (a “tag”) can be adjoined to the 5’ end of PCR primers. While these tags allow multiple samples to be pooled for simultaneous (multiplex) sequencing, they can introduce sequence-specific bias by changing primer binding strength (Berry et al. 2011). In effect, these additions simply lengthen the primer sequence.
- High throughput sequencing platforms are thought to be relatively free from sequence-specific biases, though low nucleotide diversity can degrade sequence quality (Fadrosh et al. 2014). Further, the bioinformatic protocols used to process raw sequence data can influence the inferred number of reads for a given taxon (Schloss et al. 2011).
- Lastly, the taxonomic information DNA provides varies among loci, taxa, and environments (Soergel et al. 2012), and nucleotide sequence repositories (e.g. Genbank) are incomplete and

Accepted Article

This article is protected by copyright. All rights reserved.

both geographically and taxonomically biased (Puillandre et al. 2009; Hijmans et al. 2000), limiting our ability to confidently connect identified DNA sequences with specific taxa.

The above list is not a complete set of hurdles faced by eDNA methods and we expect additional challenges will arise in the future. However, the model structure and logical process of dividing the production of eDNA into conditionally independent processes is general and broadly applicable to eDNA problems.

Statistical methods for community eDNA

As discussed above, methods for eDNA are not sufficiently well developed at present to make full inference about density or biomass in an ecological community from eDNA. Similar challenges confront estimation of density and biomass based on traditional sampling methods (Burnham et al. 1980, Hankin and Reeves 1988, Kéry and Royle 2010), but do not prevent researchers from making the best approximations possible given existing knowledge and data. In this section we provide a statistical framework for estimating the final term in eq. 3, $[\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{\xi}]$**,** in a community context. Once we have an estimate of ***Y*****,** if we can assume that the transitions from ***Y*** all the way to ***B*** do not have taxon-specific biases, our approach allows statistically-justified inferences about the relative abundance of taxa within a sampled community. As the processes related to sampling eDNA become increasingly well understood, the other three terms in eq. 3 can be modeled using a logic similar to the one detailed below.

For a sample of seawater that has been filtered, has had its total DNA extracted, amplified by PCR, and has been processed by a high-throughput sequencer, our empirical observations will be counts of unique DNA sequences. DNA sequences may be classified into types on the basis of their similarity with respect to a user-specified threshold. These are most

This article is protected by copyright. All rights reserved.

Accepted Article

often referred to as operational taxonomic units (OTUs), and hereafter we refer to them as OTUs. For simplicity, we initially treat each unique DNA sequence observed as an OTU, and later discuss how to combine distinct OTUs into groups. The results of a single sequencing run can be written as $\boldsymbol{Z}$=$\boldsymbol{z}$, where $\boldsymbol{z}$ is a realization of the random variable $\boldsymbol{Z}$ and is vector of length *I*. Each entry in the vector, $\boldsymbol{z}_i$, then contains the counts of the i^{th} OTU.

Using Bayes' theorem, we write the posterior probability of $\boldsymbol{Y}$, given our observations and parameters as proportional to the likelihood of the observations, $[\boldsymbol{Z}=\boldsymbol{z}|\boldsymbol{Y},\boldsymbol{\xi}]$, and the prior probability of the parameters $[\boldsymbol{\xi}]$**,**

$$[\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z},\boldsymbol{\xi}] \propto [\boldsymbol{Z}=\boldsymbol{z}|\boldsymbol{Y},\boldsymbol{\xi}][\boldsymbol{\xi}] \qquad (4)$$

A logical sampling model for counts with many possible categories is a multinomial model. We replace the general parameter notation $\boldsymbol{\xi}$ with $\boldsymbol{\pi}=\{\pi_1,\pi_2,\ldots\pi_I\}$ which represents the proportion of each OTU sequence present in the collected sample. Then we can write the likelihood as

$$[\boldsymbol{Z}=\boldsymbol{z}|\boldsymbol{Y},\boldsymbol{\pi}]\sim Multinomial(\boldsymbol{\pi},n) \qquad (5)$$

where n is the total number of DNA sequences observed. With a single sequencing run we have single set of observed counts, $\boldsymbol{z}$. However if we have *M* total observations of DNA sequences from a single DNA extraction – potentially from multiple independent PCR reactions or sequencing runs – we have,

$$[\boldsymbol{Z}=\boldsymbol{z}_1,\ldots,\boldsymbol{z}_M|\boldsymbol{Y},\boldsymbol{\pi}]\sim Multinomial(\boldsymbol{\pi},n_1,n_2,\ldots,n_M) \qquad (6)$$

This equation states that each $\boldsymbol{z}$ is a sample from a shared process (i.e. there is a single true proportion of DNA from each taxon in the eDNA sample and we have *M* observations of this process). Variation among the observations of $\boldsymbol{z}$ can be attributed to stochastic processes occurring during PCR and sequencing, and the model described here can be generalized to include these effects as individually modeled parameters if desired.

This article is protected by copyright. All rights reserved.

Accepted Article

Because a multinomial distribution can be written as a combination of independent Poisson distributions (the multinomial-Poisson transformation; Baker 1994), it is convenient to write the number of sequenced DNA fragments observed for each OTU as an independent Poisson random variable,

$$\begin{gathered} z_{im} \sim Poisson(e^{\lambda_i}) \\ \lambda_i = \beta_i \end{gathered} \qquad (7)$$

Here, β_i, the OTU-specific fixed effects, and λ_i are identical, but we use this notation to allow later elaboration in circumstances where additional processes are thought to influence λ_i. The proportion of DNA associated with each OTU can then be found by calculating

$$\pi_i = \frac{e^{\beta_i}}{\sum_i e^{\beta_i}} \qquad (8)$$

Note that Eq. 6 provides identical inference to eqs. 7 and 8 (Baker 1994).

The model formulation in eq. 7 assumes that each observed DNA fragment is sampled independently from a multinomial distribution. Due to the compounding process of sequential amplification in PCR, counts of DNA sequences from a sequencer are not truly independent observations of the extracted DNA. One method to deal with such non-independence is to allow for overdispersion in the Poisson parameter λ. With $m = 1,2,\ldots,M$ replicate observations, we can write the observed species counts as an over-dispersed Poisson and estimate the amount of over-dispersion, σ^2,

$$\begin{gathered} z_{im} \sim Poisson(e^{\lambda_{im}}) \\ \lambda_{im} = \beta_i + \eta_{im} \\ \eta_{im} \sim N(0, \sigma^2) \end{gathered} \qquad (9)$$

This is a simple random effects model, but one that allows great flexibility in modeling count data. Note that in the case that only a single OTU is present, eq. 9 simplifies to a log-linear model of the DNA counts and thus the single OTU version of this model is appropriate for qPCR

This article is protected by copyright. All rights reserved.

Accepted Article

data. When we observe more than one OTU, we can still produce estimates of the proportion of DNA from each taxa across all of our observations (eq. 8). After specifying prior parameters, we can use standard Bayesian Markov chain Monte Carlo (MCMC) methods to estimate the model and provide uncertainty bounds (Gelman et al. 2003). Likelihood optimization methods are also available. A further benefit of the structure is the possibility of multiple random effects that can represent multiple sources of variation in the observed counts. We present a more complicated example in the online supplement. We note that the above model is similar to other models for sequencing data proposed in a different context for other applications (Love et al. 2014).

Addressing primer bias: a framework and a simulated example

Equation 9 implicitly makes assumptions that eDNA data almost certainly violate. Importantly, eq. 9 assumes all OTUs present in the DNA extraction will be amplified equally well by PCR, and will subsequently appear in the count data emerging from the sequencer, yet PCR primers are intentionally designed to amplify specific taxonomic groups (e.g. vertebrates) to the exclusion of others (e.g., Riaz et al. 2011). Even within a target group of taxa, intra-group genetic variability in the primer binding site can cause variation in template-primer mismatch, resulting in unequal amplification among templates and thus bias in the observed sequences (e.g. Hong et al. 2009). Estimating the extent of amplification bias due to this interaction requires detailed information about both the primer set and the template (target) sequence for all taxa of interest. Generally, a way to incorporate a series of covariates—such as would describe these OTU-specific effects—is to construct a matrix of covariates, $\boldsymbol{H}$**,** and estimated coefficients, $\boldsymbol{\gamma}$, given available information about primer mismatches with existing sequence data from target

This article is protected by copyright. All rights reserved.

taxa. Accordingly, the second line of eq. 9 can be modified to accommodate variation in primer specificity to become:

$$\lambda_{im} = \beta_i + \boldsymbol{\gamma}\boldsymbol{H}_{im} + \eta_{im} \qquad (10)$$

where $\boldsymbol{\gamma}$ defines how covariates shared across taxa (e.g. the quality of match between the primer and taxa DNA) will affect the observed number of DNA sequences for each taxon. Also, note that the researcher-specified design matrix $\boldsymbol{H}$ includes the subscript *m*. This indicates multiple PCR or sequencing runs conducted using distinct methods on a single sample can be used jointly to improve the reconstruction of the ecological community of interest. For example, if two or more independent analyses were carried out on the same DNA extraction—such as in the case of multi-locus eDNA studies—the results could be formally combined into a single analysis. Furthermore, such methodological variation will help inform how changing primer specificity, the PCR reaction parameters, or other methods affect the inference about the proportion of DNA associated with each OTU. We illustrate an application of these methods below in "*Understanding marine invertebrate communities using eDNA*".

To illustrate the potential consequences of the effect of primer-template mismatch on estimates of OTU composition, we simulated small changes to the quality of primer match and used estimates of γ to show how they affected estimates in a simple three-species community (Fig. 2, supplementary materials). Simulations show that a change in primer-template match of as little as 5% (e.g. a 3 base-pair difference between a 60 bp long template and the combined forward and reverse primer) can change estimates of relative abundance (Fig. 2). The most important point of Fig. 2 is that because the estimates are relative proportions that must sum to one, if one taxon has a biased estimate, all of the other taxa's estimates are biased as well. A consequence of this observation is that analyzing data derived from multi-species primers on a

Accepted Article

This article is protected by copyright. All rights reserved.

species-by-species basis (i.e. treating the number of reads for each taxa independently in later analyses) is likely to decrease statistical precision and introduce bias in the relationship between the number of reads and virtually any other variable.

Estimating the absolute concentration of DNA in an extraction

Thus far, we have not provided direct estimates of the concentration of template DNA in the sample, $\boldsymbol{Y}$, but only estimates of the proportional abundance of each OTU, $\boldsymbol{\pi}$**.** To generate estimates of DNA concentration, we need to incorporate additional information about the absolute abundance of DNA from at least some of the OTUs to scale the proportional abundance to true abundance. We can use the posterior estimates of proportional abundance $\boldsymbol{\pi}$ in combination with posterior estimates of the density of DNA from a single OTU, ω_1, to scale the proportions to DNA densities for all OTUs. Current methods using qPCR are adept at producing estimates of ω_1 (Jerde et al. 2011, Lodge et al. 2012, Takahara et al. 2012). If we assume that ω_1 and $\boldsymbol{\pi}$ are derived from independent methods, we can use draws from the posterior distributions of each to derive the posterior distribution of $\boldsymbol{Y}$**.** For the j^{th} OTU and g^{th} draw from the posterior distribution, we have

$$Y_j^{(g)} = \omega_1^{(g)} \left(\frac{\pi_j^{(g)}}{\pi_1^{(g)}} \right) \qquad (11)$$

After calculating $\boldsymbol{Y}$ for a large number of posterior draws, we can summarize $\boldsymbol{Y}$ using standard descriptors (mean, standard deviation, etc.). This method is appealing because it reflects the uncertainty in both $\boldsymbol{\pi}$ and the concentration of DNA derived from qPCR. It also shows how qPCR and sequencing approaches are complementary data types that can be combined and re-emphasizes how the structure presented in Figure 1 is applicable to a wide variety of eDNA methods. We highlight the utility of this two-pronged validation method for future applications.

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

Detection probabilities and power analysis

A trade-off between detection probability for any given taxon and breadth of the community observed is common to surveys using both eDNA and non-molecular (i.e., traditional) methods. In many eDNA applications, the risk of false-negative detections (in which a taxon is present, but not detected) is one of the most pressing issues (Yoccoz 2012, Yu et al. 2012, Ji et al. 2014,). Conveniently, the model outlined in eqs. 9 and 10 provides a method for determining the thresholds for detection. However, because the PCR primers for community eDNA analyses will almost never be strictly taxon-specific, the power analysis cannot be determined on a single-taxon basis but must always be phrased in terms of a larger DNA community that is "observed" by a given PCR protocol.

The relative abundance of an arbitrary OTU, taxon "A", can be fully defined by four quantities: the true relative proportion of DNA from OTU A in the sample π_A; the estimated effect of covariates for that OTU, $\boldsymbol{\gamma H_A}$; the total number of DNA sequences observed, *n;* and the stochasticity in the PCR and sequencing process, σ^2. Because for the observed data, $n = \sum_i e^{\lambda_i}$ (eq. 8), we can combine eq. 8 and 10 and use the properties of the log-normal distribution to show that for any true value of π_A, the median value of λ_A, λ_A^*, will be

$$\lambda_A^* = \log(\pi_A) + \log(n) + \boldsymbol{\gamma H_A} \qquad (12)$$

Using the probability mass function of the Poisson distribution, the probability that the observed number of DNA sequences for OTU *A* will exceed 0 at λ_A^* is,

$$p(z_A > 0) = 1 - e^{\lambda_A^*} \qquad (13)$$

In this way, the detection probability can be approximated for a given primer, the number of DNA sequences observed, and DNA community. This type of power analysis based on the

This article is protected by copyright. All rights reserved.

median estimate is likely sufficient for most applications, but it is important to acknowledge that this approach ignores variability in the PCR process (σ^2) and uncertainty in the estimate of $\boldsymbol{\gamma}$. However, simulation approaches could incorporate this variability if desired. Importantly, eqs. 12 and 13 make explicit that analytical approaches based on the occurrence data (Yu et al. 2012, Ji et al. 2014) are special cases of multi-taxa count data. In its simplest form, occurrence data is simply the count data for each OTU converted into two classes: $z_i = 0$ and $z_i > 0$. Other investigators have suggested that OTUs below a certain threshold abundance should be excluded (e.g. OTUs below 0.005% of the total number of DNA reads is recommended by Bokulich et al. 2013). Regardless of the exact cutoff used, this section demonstrates that the same biases that plague estimating abundance from eDNA will also plague estimations of occurrence – though signatures of bias will be more difficult to detect and estimate using occurrence data.

We illustrate power curves in Fig. 3 to provide a graphical method for understanding the detection probability of a taxon for a given primer, extracted DNA, and number of DNA reads. Specifically, we compare three values of a single covariate representing the match between the primer and taxon *A*'s DNA. H_A = 0 represents the average match between the primer and the taxa observed in the sample, while H_A = -0.15 corresponds to *A* having a 15% better match to primer than average and H_A = 0.15 corresponds to *A* having a 15% worse match to primer than average (e.g. for a 20 basepair primer, 15% corresponds to a change of 3 basepair matches between primer and template). For all three simulations, we used a slope parameter that reflect real-world estimates of primer bias discussed below in "Understanding marine invertebrate communities using eDNA" $(\gamma = -14)$. An important result of Fig. 3 is that even when a taxon is present in a sample, it may not be observed in the DNA counts emerging from the sequencer. The probability of observing at least one instance of taxon A is affected both by its true

This article is protected by copyright. All rights reserved.

abundance (relative to other species amplified by the PCR product) and the match between the DNA sequence and the PCR primer used.

Eq. 12 and Fig. 3 suggest that there are several intuitive and non-mutually exclusive methods for increasing detection probability of a particular taxon: 1) increase the number of sequences observed for each PCR (increase *n*); 2) decrease the number of taxa amplified by the primer (decrease *I* and thereby increase the relative abundance of the OTU of interest, π_A); 3) improve the efficiency of the primer for taxon A relative to other taxon in the DNA community (i.e. modify $\boldsymbol{H_A}$). In practice, a PCR primer that more closely matches a particular taxon will likely contribute to both point 2 and 3. However, increased primer specificity will always reduce the diversity of taxa detected in a single sequencing run. Both highly specific and more general primers have important real world applications (Simmons et al. *In Press*).

Combining unique DNA sequences into biologically meaningful groups

Genetic variation among individuals both within and across taxa can result in two problematic scenarios: 1) high diversity within a taxon will result in it being represented by more than one OTU in the sequence data or 2) low diversity across taxa will result in many taxa being represented by a single OTU. An ideal PCR primer would target a locus with high inter-taxon diversity and low intra-taxon diversity. Unfortunately, we know of no such locus that can be used for a broad swath of taxa. For the case where a single taxon is represented by multiple OTUs, we describe two approaches for obtaining abundance estimates.

The first is to estimate the model treating each OTU separately (eq. 12), and combine the output of the estimation procedure. Because each iteration of a Markov chain provides a draw from the posterior distribution of the parameters, the draws can simply be added together for the

This article is protected by copyright. All rights reserved.

Accepted Article

OTUs of interest, and the proportion of the resulting taxon recalculated (Shelton et al. 2012). To provide a concrete, but fictitious, example, suppose that OTU *A* and OTU *B* were both observed in a sequencing run. Both OTUs are subsequently determined to represent unique sequences from woolley mammoth *(Mammuthus primigenius)* and need to be combined to provide an estimate of the total mammoth present in the extracted DNA sample. After estimating a Poisson model (e.g. eq. 10) we can simply add the two estimated parameters for OTU *A* and OTU *B* (β_A and β_B, respectively) such that $\beta_{mammoth} = \beta_A + \beta_B$ for each MCMC iteration. The proportion of DNA attributable to mammoth would then be $\pi_{mammoth} = \frac{e^{\beta_{mammoth}}}{\sum_i e^{\beta_i}}$. Using draws from the posterior distribution maintains the correlation structure and uncertainty bounds of the proportional occurrence. However, this approach has the downside of requiring parameter estimates and the collection of covariates to populate ***H*** for each OTU, slowing computation speed if there are large numbers of OTUs.

The second option is to group the OTUs into broader taxonomic groups before they are included as input data for the model estimation. While the choice of method for clustering sequence data into OTU counts is of general concern (Edgar et al. 2011, Yu et al. 2012), this approach also requires that all OTUs within a group be assumed to have shared covariates related to PCR. Continuing our previous mammoth example, the primer-template mismatch might differ between OTU *A* and OTU *B*, and yet if their counts were to be combined, information about their distinct matching characteristics could not be directly incorporated in the model. A summary statistic such as the median dissimilarity would have to be used instead. Depending on the details of the primers and match quality, such averaging across covariates may or may not substantially influence the result. Given these considerations, we advocate the first approach of combining

This article is protected by copyright. All rights reserved.

Accepted Article

taxa after model estimation, unless speed is favored over accuracy or researchers are sufficiently confident that grouped taxa do not differ in PCR or sequencing efficiency.

Understanding marine invertebrate communities using eDNA

To illustrate the utility of our statistical framework, we apply the above methods to eDNA isolated, amplified, and sequenced from eleven, 1-L seawater samples collected from a single location in Puget Sound, WA on June 26, 27, and 29, 2014 (Carkeek Park, Seattle, WA, USA; 47°42'40.44"N, 122°22'20.10"W). Because we use this empirical dataset here only to illustrate the application of statistical methods to counts of DNA sequences emerging from a high throughput sequencer, we only outline the methods that affect the statistical estimation. We provide detailed molecular protocols in the online supplement for interested readers.

Summary of molecular methods

To test the effect of primer mismatch on template-specific PCR efficiency, we amplified each environmental sample using two different sets of primers, which in each direction shared a common core 22bp region targeting the 16S region of the mitochondrion, but differed by an index sequence on the 5’ end (see Table S1 for the primer sequences used). These index sequences have been demonstrated to cause differential amplification efficiency among template DNA in a mixed-template PCR (Berry et al. 2011), and thus provide an opportunity to test the efficacy of our framework for estimating biomass and uncertainty in the face of bias. PCR, library preparation, sequencing, and bioinformatics protocols are described in the supplementary material.

The experimental design yielded sequence data from six PCR products per environmental sample: three sequencing replicates arising from each of two distinct primer sets. In total, we

This article is protected by copyright. All rights reserved.

Accepted Article

observed over 10.5 million individual DNA reads representing 27,973 unique OTUs. For the purpose of this example, we model only 9 of the most common OTUs and focus on estimating the proportional DNA contribution for these 9 OTUs and a tenth "Other" category which encompasses all remaining OTUs. We investigate only 10 OTUs for illustration purposes, though this approach is directly applicable to a much larger set of OTUs. We present the raw data and models for estimating these models for these nine OTUs in the supplementary materials.

Statistical modeling of OTU counts

To estimate the proportion of each of these 9 OTUs on each sampling occasion, we use a version of eq. 10 that adds a subscript *t* corresponding to each sample time and includes *m* observed DNA replicates for each time. Then the full model is

$$\begin{gathered} z_{itm} \sim Poisson\left(e^{\lambda_{itm}}\right) \\ \lambda_{itm} = \beta_{it} + \boldsymbol{\gamma H_{itm}} + \eta_{itm} \\ \eta_{itm} \sim N\left(0, \sigma^2\right) \end{gathered} \tag{14}$$

Again, β_{it} indicates the count of OTU *i* at time *t*, $\boldsymbol{\gamma H_{itm}}$ controls the fixed effect of PCR and sequencing bias on the observed number of OTU counts for each replicate, with $\boldsymbol{\gamma}$ estimated regression coefficients and the covariate matrix $\boldsymbol{H_{itm}}$ supplied by the investigator on the basis of available information about target-taxon sequences in the primer region. Finally, η_{itm} provides for additional error not accounted for by either the fixed taxon effect β_{it}or the other fixed effects. While it is possible to include a large variety of potential covariates in $\boldsymbol{\gamma H_{itm}}$ for illustration purposes we include only a single covariate, the total genetic distance between the OTUs' primer binding sites and the primers, γ, at both forward and reverse priming sites. Thus $\boldsymbol{H}$ is a design matrix with a single column corresponding to the proportion of nucleotide mismatches between the primers and each template (OTU primer binding site). A value of 0 would indicate no difference between the primer and the template, while 0.10 would indicate 10% of base pairs do

This article is protected by copyright. All rights reserved.

Accepted Article

not match between the primer and the OTU. Distance calculations were performed using the function dist.dna in the R package ape (Paradis et al. 2004). To derive estimates of the design matrix ***H*** we assessed the quality of match between the primer and each taxon's DNA. For the nine focal OTUs in the dataset, we performed a BLAST search of NCBI's nucleotide database (GenBank) to identify the likely sequence of the primer binding sites given existing sequence information for taxa in GenBank matching the OTU sequences (see below). We centered the covariate values in ***H*** before analysis by subtracting each value by the average across all OTU-primer pairs. The process of centering makes β_{it} the intercept for each OTU in this generalized linear model. We assumed the "Other" category had a covariate value of 0, (i.e. $H_{Other,t,m} = 0$) corresponding to the average amplification value of the "Other" category. Centering the covariates also means that when we calculate the proportional contribution of each OTU, we can calculate the proportion of each OTU in the sample as $\pi_{it} = \frac{e^{\beta_{it}}}{\sum_i e^{\beta_{it}}}$. This produces estimates of proportional composition of each OTU at a standardized match between the primer and substrate for all OTUs.

We estimated eq. 14, using Just Another Gibbs Sampler (JAGS; Plummer 2003) implemented in R (R Core Team 2014) using the R2jags package (Su and Yajima 2014). We used non-informative prior distributions for each parameter. Specifically we let $\gamma \sim Normal(0,1000)$, $\beta\_ \sim Normal(0,1000)$, and $\sigma^2 \sim Uniform(0,1000)$. We ran three replicate MCMC chains using a 100,000 iteration burn-in and 10,000 monitoring iterations. We confirmed appropriate model mixing and convergence using visual inspection of trace plots and Gelman-Rubin diagnostics as implemented by the R package "coda" (Plummer et al. 2006).

This article is protected by copyright. All rights reserved.

Accepted Article

Results

Using eq. 14, we estimated the proportional composition for nine focal OTUs and the "Other" category for all eleven time periods (Fig. 4, Fig. 5). Our model estimated a large amount of overdispersion in the observed count data ($\sigma^2 = 8.34[0.68]$; posterior mean[sd]) indicating that there remains a substantial effect of unknown and unmodeled factors on variation among samples. The large estimated overdispersion translates into large uncertainty in the estimated proportional composition (Fig. 4). Our estimates are statistically well-justified and reflect the uncertainty present in our observations, but suggest that methodological improvements will be required to provide more precise estimates of the marine community. Across all times, OTUs 3, 5, and 7 were particularly frequent. Both OTU 3 and 7 correspond to the mussel, *Mytilus trossulus*, while OTU 5 corresponds to acorn barnacles (suborder Balanomorpha; likely *Balanus glandula*), both of which are among the most commonly observed species at our study site. We found no dramatic patterns of OTU relative abundance over time or with respect to an important covariate, tidal height (Fig. 5). However, the large degree of uncertainty limits our power to detect strong effects of time or environmental factors.

Among our nine focal OTUs—which, again, represent sequences amplified and recovered from environmental samples—the variance in primer-template mismatch was substantial. Across all primer-template pairs the mean proportional mismatch was 0.193 (range: 0.11-0.28), indicating that, on average 10.81 out of a total 56 base pairs were mismatched. We estimated, as expected, that the effect of decreasing match between the primer and substrate was strongly negative, $\gamma = -14.3[6.11]$(posterior mean[sd]) indicating OTUs with a poor match between the primer binding site and primer were underrepresented in the observed DNA counts. Our estimated effect of primer quality is similar to experimental results exploring the effect of

This article is protected by copyright. All rights reserved.

primer mismatch on preferential PCR amplification (Polz 1998, Sipos 2007, Wright et al. 2014). We emphasize that there are a great many possible other covariates that could be used in this type of analysis.

Discussion and conclusions

eDNA is an exciting emerging method for describing ecological communities. Given the enormous potential for eDNA applications in the environmental sciences, recent reviews of eDNA methods have stressed the need for improved molecular and statistical techniques for eDNA (Yu et al. 2012, Yoccoz et al. 2012, Schmidt et al. 2013, Ji et al. 2014). Conceptually, the challenges posed by eDNA are largely analogous to those faced by traditional sampling techniques (Fig. 1). Both conventional and eDNA sampling ultimately attempt to make inferences about the same quantity: the biomass or density of each species in the environment. It should also be clear that traditional sampling methods suffer from a parallel set of sampling problems to eDNA and, as noted earlier, our current inability to estimate abundance or biomass from eDNA samples alone is not a fatal flaw for eDNA data. A specific topic that deserves special consideration in future work is understanding the spatial and temporal spread of eDNA under natural conditions and how the scale of inference from eDNA sampling matches (or, potentially, does not match) the spatial and temporal inference available from traditional sampling methods.

While we have framed our analysis in terms of biomass, we note that an equivalent structure is necessary for estimation of count data and for deriving most community metrics of interest as well. Estimated species richness is the number of species with biomass greater than 0 while Shannon diversity is species richness weighted by the relative biomass (or count) of each

This article is protected by copyright. All rights reserved.

Accepted Article

species. Both richness and Shannon diversity – and indeed virtually all community and diversity metrics – are directly derived from estimates of occurrence and abundance of individual species. Thus this framework provides a pathway for investigating communities as well as individual taxa.

In closing we offer a few recommendations to ensure that eDNA study designs—and the resulting datasets—are adequate to develop a meaningful estimate of the target biological community structure.

Foremost, it should be clear from the framework we discuss here that sample replication (in space, time, laboratory treatment, etc.) is critical to partitioning variance among steps in the eDNA analytical chain. Because real-world constraints on time and funding generally prevent replication at every step, we emphasize that replication is most important at the step or steps that are likely to introduce the greatest amount of variance or where the variance attributable to that step is of special interest. For example, if one has data demonstrating that eDNA capture, extraction, and sequencing are likely to introduce little systematic bias, but that PCR primer choice has an unknown and potentially large effect, PCR is the most important target for replication and independent analysis. Samples treated separately can then subsequently be combined using hierarchical models, where this would provide analytical benefit (see online supplement). Note additionally that we advocate avoiding pooling samples and then running analyses on the pooled output whenever possible; there is information in the variability among replicated outputs of molecular methods.

Second, because taxa are not equally abundant in a sampled environment, and because taxa are not equally likely to amplify with a given set of PCR primers, eDNA community surveys are necessarily an uneven reflection of taxa present, even for a specifically targeted

Accepted Article

This article is protected by copyright. All rights reserved.

groups. The same issues arise with traditional sampling methods, as alternative survey methods have different but non-negligible selectivity issues (Beverton and Holt 1957 section 8, Arreguín-Sánchez 1996, Venebles and Dichmont 2004).

The methods we present for community eDNA data offer the ability to correct for attributable biases and to be statistically honest about biases and variability that we do not understand. However, real differences in DNA abundance and susceptibility to amplification mean that for any given set of PCR primers there is a limited set of taxa that can successfully be detected. This observation gives rise to three recommendations:

1. Using multiple markers offers the chance to broaden the scope of an eDNA survey and to generate mutually reinforcing datasets that might be combined in the framework we present here (Evans et al. 2015).
2. Community surveys that focus on the most common sequences generated—rather than on the rare sequence "tails"—are more likely to be repeatable and robust to statistical inference. At the same time, we acknowledge that some analyses – particularly those focused on measures of biodiversity (e.g. Ji et al. 2014) - are intrinsically interested in rare taxa. We think an increased focus on understanding how the probability of detection may affect diversity estimates is an important area for further research (Fig. 2; Schmidt et al. 2013).
3. Finally, a focus on the most common species (or most common sequences) found in an environment has implications for primer design. Rather than accepting a very broad set of sequence constraints on primer design (e.g., all metazoans), ensuring that primers are likely to be good matches for the few dozen most common target species in the sampled area is likely to yield a better range of acceptable primer sequences. Increased specificity is more likely to lead to the intended results of a community eDNA survey. Again, this approach is

Accepted Article

This article is protected by copyright. All rights reserved.

appropriate only when the interest is focused on relatively common species, not on rare or unknown taxa in the community.

As we have suggested throughout this paper, we believe there is ample room for cross-pollination between eDNA, both qPCR and sequencing based, and traditional sampling approaches. Notably, the conceptual framework we outline suggests that it is possible to construct models that jointly model data from traditional and eDNA sampling to draw inference about natural populations. We also expect that methodological biases inherent to eDNA and traditional sampling may often produce complementary, rather than overlapping, estimates of community composition. Regardless, here we have shown how to start toward this ultimate goal by providing a framework and detailed statistical models for a particularly challenging aspect of eDNA work—calculating the relative abundance of DNA from multi-species primers while accounting for variation in PCR. However, multiple elements of the eDNA processing chain remain poorly described from a quantitative perspective, and as future work clarifies biases introduced at each experimental step, our framework provides a means of using such emerging information to improve quantitative estimates of community biomass from eDNA.

Acknowledgements

This work was supported in part by a grant to RPK from the David and Lucile Packard Foundation. Thanks also to the Helen R. Whiteley Center at Friday Harbor Laboratories for supporting the writing workshop that substantially advanced this product. We thank E. Buckner, E. Garrison, M. Klein, and A. Wong for help with field collections and Seattle Parks and Recreation for access to Carkeek Park. We thank P. Levin, A. Stier, B. Feist, K. Marshall, and S.

Accepted Article

This article is protected by copyright. All rights reserved.

Hennessey for comments on earlier versions of this manuscript. K. Deiner and three anonymous reviewers improved this manuscript.

Literature Cited

Arreguín-Sánchez, F. 1996. Catchability: a key parameter for fish stock assessment. Reviews in Fish Biology and Fisheries 6:221–242.

Baker, S.G. 1994. The multinomial-Poisson transformation. The Statistician 43:495.

Barnes, M.A., C.R. Turner, C.L. Jerde, M.A. Renshaw, W.L. Chadderton, and D.M. Lodge. 2014. Environmental conditions influence eDNA persistence in aquatic systems. Environmental Science and Technology 48:1819–1827.

Berry, D., K.B. Mahfoudh, M. Wagner, and A. Loy. 2011. Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. Applied and Environmental Microbiology 77:7846–7849.

Beverton, R. J. H., and S. J. Holt. 1957. On the dynamics of exploited fish populations. Chapman and Hall.

Bokulich, N.A., S. Subramanian, J.J. Faith, D. Gevers, J.I. Gordon, R. Knight, D.A. Mills, and J.G. Caporaso. 2013. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nature Methods 10:57–59.

Branch, T.A., R. Watson, E.A. Fulton, S. Jennings, C.R. McGilliard, G.T. Pablico, D. Ricard, S.R. Tracey. 2010. The trophic fingerprint of marine fisheries. Nature 468:431-435.

Burnham, K.P., D.R. Anderson, and J.L. Laake. 1980. Estimation of density from line transect sampling of biological populations. Wildlife Monographs 3–202.

Camacho C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T.L. Madden.

Accepted Article

This article is protected by copyright. All rights reserved.

2009. BLAST+: architecture and applications. BMC bioinformatics 10:421.

Carrigg, C., O. Rice, S. Kavanagh, G. Collins, and V. O'Flaherty. 2007. DNA extraction method affects microbial community profiles from soils and sediment. Applied Microbiology and Biotechnology **77**:955–964.

Clark, J.S. 2007. Models for ecological data. Princeton, Princeton, NJ.

Cochran, W.G. 1977. Sampling techniques. John Wiley and Sons Inc.

Cotter, A.J.R., and G.M. Pilling. 2007. Landings, logbooks and observer surveys: improving the protocols for sampling commercial fisheries. Fish and Fisheries 8:123–152.

Cressie, N., and C.K. Wikle. 2011. Statistics for spatio-temporal data. John Wiley and Sons Inc.

de Vargas, C., S. Audic, N. Henry, J. Decelle, F. Mahé, R. Logares, et al. 2015. Eukaryotic plankton diversity in the sunlit ocean. *Science, 348:* 1261605.

Deiner, K. and F. Altermatt F. 2014. Transport distance of invertebrate environmental DNA in a natural river. PLoS ONE 9: e88786. doi:10.1371/journal.pone.0088786

Deiner, K., J.-C. Walser, E. Mächler, and F. Altermatt. 2015. Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. Biological Conservation **183**:53–63.

Edgar R.C. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460-2461.

Edgar, R.C., B.J. Haas, J.C. Clemente, C. Quince, and R. Knight. 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27:2194–2200.

Elith, J. and J.R. Leathwick. 2009. Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics 40:677-697.

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

Evans, N.T., B.P. Olds, M.A. Renshaw, C.R. Turner, Y. Li, C.J. Jerde, A.R. Mahon, M.E. Pfrender, G.A. Lamberti, and D.M. Lodge. 2015. Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. Molecular Ecology Resources. doi: 10.1111/1755-0998.12433

Fadrosh, D.W., B. Ma, P. Gajer, N. Sengamalay, S. Ott, R.M. Brotman, and J. Ravel. 2014. An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. Microbiome 2:1–7.

Feinstein, L.M., W.J. Sul, and C.B. Blackwood. 2009 Assessment of bias associated with incomplete extraction of microbial DNA from soil. Applied Environmental Microbiology. 75:5428–5433.

Hankin, D.G. and G.H. Reeves. 1988. Estimating total fish abundance and total habitat area in small streams based on visual estimation methods. Canadian Journal of Fisheries and Aquatic Sciences 45:834–844.

Hijmans, R.J., K.A. Garrett, Z. Huamán, D.P. Zhang, M. Schreuder, and M. Bonierbale. 2000. Assessing the geographic representativeness of genebank collections: the case of Bolivian wild potatoes. Conservation Biology 14:1755–1765.

Iversen, L.L., J. Kielgast, and K. Sand-Jensen. *In press*. Monitoring of animal abundance by environmental DNA- An increasingly obscure perspective: A reply to Klymus et al., 2015. Biological Conservation. http://dx.doi.org/10.1016/j.biocon.2015.09.024

Ficetola, G.F, J. Pansu, A. Bonin, E. Coissac, C. Giguet-Covex, M. De Barba, L. Gielly, C.M. Lopes, F. Boyer, F. Pompanon, G. Rayé and P. Taberlet. 2015. Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. Molecular Ecology Resources. 15: 543–556

This article is protected by copyright. All rights reserved.

Accepted Article

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. 2003. Bayesian data Analysis. Second edition. Chapman and Hall/CRC.

Hong S., J. Bunge, C. Leslin, S. Jeon, and S.S. Epstein. 2009. Polymerase chain reaction primers miss half of rRNA microbial diversity. ISME 3: 1365–73.

Jannot, J.E. and D.S. Holland 2013. Identifying ecological and fishing drivers of bycatch in a U.S. groundfish fishery. Ecological Applications 23:1645–1658.

Jerde, C.L. A.R. Mahon, W.L. Chadderton, and D.M. Lodge 2011. 'Sight-unseen' detection of rare aquatic species using environmental DNA. Conservation Letters 4:150-157.

Ji, Y., L. Ashton, S.M. Pedley, D.P. Edwards, Y. Tang, A. Nakamura, R. Kitching, P.M. Dolman, P. Woodcock, F.A. Edwards, T.H. Larsen, W.W. Hsu, S. Benedick, K.C. Hamer, D.S.Wilcove, C. Bruce, X. Wang, T. Levi, M. Lott, B.C. Emerson, and D.W. Yu. 2013. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecology Letters 16:1245-1257.

Kéry, M., and J.A. Royle. 2010. Hierarchical modelling and estimation of abundance and population trends in metapopulation designs. Journal of Animal Ecology 79:453–461.

Kelly, R.P., J.A. Port, K.M. Yamahara, and L.B. Crowder. 2014 Using Environmental DNA to Census Marine Fishes in a Large Mesocosm. PLoS ONE 9: e86175

Kimura, D.K. and H.H. Zenger, Jr. 1997. Standardizing sablefish (*Anoplopoma fimbria*) long-line survey abundance indices by modeling the log-ratio of paired comparative fishing cpues. ICES Journal of Marine Science, 54:48–59.

Klymus, K., C.A. Richter, D. Chapman, and C. Paukert. 2015. Quantification of eDNA shedding rates from invasive bighead carp *Hypophthalmichthys nobilis* and silver carp *Hypophthalmichthys molitrix*. Biological Conservation 183:77–84

This article is protected by copyright. All rights reserved.

Accepted Article

Laramie, M.B., D.S. Pilliod, and C.S. Goldberg. 2015. Characterizing the distribution of an endangered salmonid using environmental DNA analysis. Biological Conservation. 183:29-37.

Lee, C.K., C.W. Herbold, S.W. Polson, K.E. Wommack, S.J. Williamson, I.R. McDonald, and S.C. Cary. 2012. Groundtruthing next-gen sequencing for microbial ecology–biases and errors in community structure estimates from PCR amplicon pyrosequencing. PloS One, **7**:e44224.

Leray, M., and N. Knowlton. 2015. DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. Proceedings of the National Academy of Sciences, 112:2076–2081.

Lodge, D.M., C.R. Turner, C.L. Jerde, M.A. Barnes, L. Chadderton, S.P. Egan, J.L. Feder, A.R. Mahon, and M.E. Pfrender. 2012. Conservation in a cup of water: estimating biodiversity and population abundance from environmental DNA. Molecular Ecology **21**: 2555–2558.

Love, M.I, W. Huber, and S. Anders. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15:550.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17:10-12.

Maruyama A., K. Nakamura, H. Yamanaka, M. Kondoh, and T. Minamoto. 2014. The release rate of environmental DNA from juvenile and adult fish. PLoS ONE 9:e114639.

Merkes C.M., S.G. McCalla, N.R. Jensen, M.P. Gaikowski, and J.J. Amberg. 2014 Persistence of DNA in carcasses, slime and avian feces may affect interpretation of environmental DNA data. PLoS ONE 9(11): e113346. doi:10.1371/journal.pone.0113346.

Nathan, L.M., M. Simmons, B.J. Wegleitner, C.L. Jerde, and A.R. Mahon. 2014. Quantifying

This article is protected by copyright. All rights reserved.

Accepted Article

environmental DNA signals for aquatic invasive species across multiple detection platforms. Environmental Science and Technology 48:12800-12806.

O'Donnell, J.L., R.P. Kelly, N. Lowell, and J.A. Port. 2015. Indexed PCR primers induce template-specific bias in large-scale DNA sequencing studies. *In revision.* PLoS One, July 2015.

Paradis E., J. Claude, and K. Strimmer. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289-290.

Pedersen, M.W., S. Overballe-Petersen, L. Ermini, C. Der Sarkissian, J. Haile, M. Hellstrom, J. Spens, P.F. Thomsen, K. Bohmann, E. Cappellini, I.B. Schnell, N.A. Wales, C. Carøe, P.F. Campos, A.M.Z. Schmidt, M.T.P. Gilbert, A.J. Hansen, L. Orlando, and E. Willerslev. 2015. Ancient and modern environmental DNA. Philosophical Transactions of the Royal Society B 370:20130383.

Pilliod, D.S., Goldberg, C. S., Arkle, R. S., and Waits, L. P. (2014). Factors influencing detection of eDNA from a stream-dwelling amphibian. Molecular Ecology Resources **14**:109–116.

Plummer, M. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2006. CODA: Convergence Diagnosis and Output Analysis for MCMC 6:7–11.

Polz, M.F., and C.M. Cavanaugh. (1998). Bias in template-to-product ratios in multitemplate PCR. Applied and Environmental Microbiology 64: 3724–3730.

Puillandre, N., E.E. Strong, P. Bouchet, M.-C. Boisselier, A. Couloux, and S. Samadi. 2009. Identifying gastropod spawn from DNA barcodes: possible but not yet practicable. Molecular Ecology Resources 9:1311–1321.

This article is protected by copyright. All rights reserved.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Renshaw, M.A, B.P. Olds, C.L. Jerde, M.M McVeigh, and D.M. Lodge. 2015. The room temperature preservation of filtered environmental DNA samples and assimilation into a phenol–chloroform–isoamyl alcohol DNA extraction. Molecular Ecology Resources 15:68-176.

Riaz T., W. Shehzad, A. Viari, F. Pompanon, P. Taberlet, E. Coissac. 2011. ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. Nucleic Acids Research 39:e145. doi:10.1093/nar/gkr732.

Royle, J.A., and J.D. Nichols. 2003. Estimating abundance from repeated presence–absence data or point counts. Ecology 84:777–790.

Schloss, P.D., D. Gevers, and S.L. Westcott. 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoSONE 6: e27310. doi:10.1371/journal.pone.0027310

Schmidt, B.R., M. Kéry, S. Ursenbacher, O.J. Hyman, and J.P. Collins. 2013. Site occupancy models in the analysis of environmental DNA presence/absence surveys: a case study of an emerging amphibian pathogen. Methods in Ecology and Evolution 4:646–653

Shelton, A.O., E.J. Dick, D.E. Pearson, S. Ralston, and M. Mangel. 2012. Estimating species composition and quantifying uncertainty in multispecies fisheries: hierarchical Bayesian models for stratified sampling protocols with missing data. Canadian Journal of Fisheries and Aquatic Sciences 69:231–246.

Simmons, M., A. Tucker, W.L. Chadderton, C.L. Jerde, A.R. Mahon. *In Press*. Active and passive environmental DNA surveillance of aquatic invasive species. Canadian Journal

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

of Fisheries and Aquatic Sciences.

Sipos, R., Székely, A.J., Palatinszky, M., Révész, S., Márialigeti, K., and Nikolausz, M. 2007. Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. FEMS Microbiology Ecology 60:341–350.

Soergel, D.A.W., N. Dey, R. Knight, and S.E. Brenner. 2012. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. The ISME Journal 6:440–1444.

Strickler, K.M., A.K. Fremier, and C.S. Goldberg. Quantifying effects of UV-B, temperature, and pH on eDNA degradation in aquatic microcosms. 2015. Biological Conservation 183: 85-92.

Su, Y.-S., and M. Yajima. 2014. R2jags: A package for running jags from R. R package version 0.04-03.

Takahara, T., T. Minamoto, H. Yamanaka, H. Doi, and Z. Kawabata. 2012. Estimation of fish biomass using environmental DNA. *PLoS ONE*, ***7:***e35868

Thomsen, P.F., J. Kielgast, L.L. Iversen, P.R. Møller, M. Rasmussen, and E. Willerslev. 2012. Detection of a diverse marine fish fauna using environmental DNA from seawater samples. PloS One, 7:e41732.

Thomsen, P.F., J. Kielgast, L.L. Iversen, C. Wiuf, M. Rasmussen, M.T.P. Gilbert, L. Orlando, and E. Willerslev. 2012. Monitoring endangered freshwater biodiversity using environmental DNA. Molecular Ecology 21: 2565-2573.

Turner, C.R., M.A. Barnes, C.C.Y. Xu, S.E. Jones, C.L. Jerde, and D.M. Lodge. 2014. Particle size distribution and optimal capture of aqueous macrobial eDNA. Methods in Ecology and Evolution **5**:676–684.

This article is protected by copyright. All rights reserved.

Accepted Article

Venables, W.N. and C.M. Dichmont. 2004. GLMs, GAMs and GLMMs: an overview of theory for applications in fisheries research. Fish. Res. 70:319–37.

Venter, J.C., K. Remington, J.F. Heidelberg, A.L. Halpern, D. Rusch, J.A. Eisen, D. Wu et al. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66-74.

Wright, E.S., L.S. Yilmaz, S. Ram, J.M. Gasser, G.W. Harrington, and D.R. Noguera. 2014. Exploiting extension bias in polymerase chain reaction to improve primer specificity in ensembles of nearly identical DNA templates. Environmental Microbiology 16:1354-1365.

Yoccoz, N.G. 2012. The future of environmental DNA in ecology. 2012. Molecular Ecology 21:2031-2038.

Yu, D.W., J. Yinqiu, B.C. Emerson, X. Wang, C. Ye, C. Yang, and Z. Ding. 2012. Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods in Ecology and Evolution 3:613–623

Zhang J.J., K. Kobert, T. Flouri, and A. Stamatakis. 2014. PEAR: a fast and accurate Illumina paired-end read merger. Bioinformatics 30:614-620.

This article is protected by copyright. All rights reserved.

Accepted Article

Figure Captions:

Figure 1: A schematic illustration of the process of sampling ecological communities using eDNA and traditional sampling methods. Boxes correspond to latent states, while arrows and greek letters represent process contributing to the transitions between states. While we present only one eDNA path and one traditional sampling path, recognize that there are many potential variations on the form of this figure depending on the details of a particular protocol.

Figure 2: The effect of primer bias on estimated species composition. In *a*, taxa A, B, and C are equally abundant but *A* and *B* had identical, average, primer-template matches ($H = 0$) while *C* had a below average match to the primer ($H = 0.05$). In panel *b*, taxa *A* and *B* are again identical, but *C* has improved primer match (H=-0.05). The dashed line shows true proportional abundance. Uncertainty bounds represent both uncertainty in the parameter γ and approximate uncertainty in the posterior distribution and covariance in the taxon-specific parameters, $\boldsymbol{\beta}$**,** derived from our example application (see supplementary materials for full simulation details). In both panels, boxplots show median, interquartile range, and 95% credible intervals and “x” denotes the mean estimate. The horizontal dashed line represents the true proportional abundance. The most important point is that because the estimates are relative proportions that must sum to one, if one taxon has a biased estimate, all of the other taxa’s estimates are biased as well.

Figure 3: The expected value of DNA reads for hypothetical taxon *A* at varying levels of true proportional abundance, π_A, for a single PCR primer and 1,000,000 total DNA reads. The three lines correspond to three possible qualities of match between primer and taxon A. If taxon A has

This article is protected by copyright. All rights reserved.

a better than average match with the primer (*H=-0.15*) it is expected to be observed more frequently than if taxon had an primer match equivalent to the other species in the sample (*H=0*) or a below average match (*H=0.15*). The right axis shows the probability that the count of taxon *A* is non-zero. See text for details. Note that if DNA is infrequent, the probability of detection can be dramatically affected by the primer-template match.

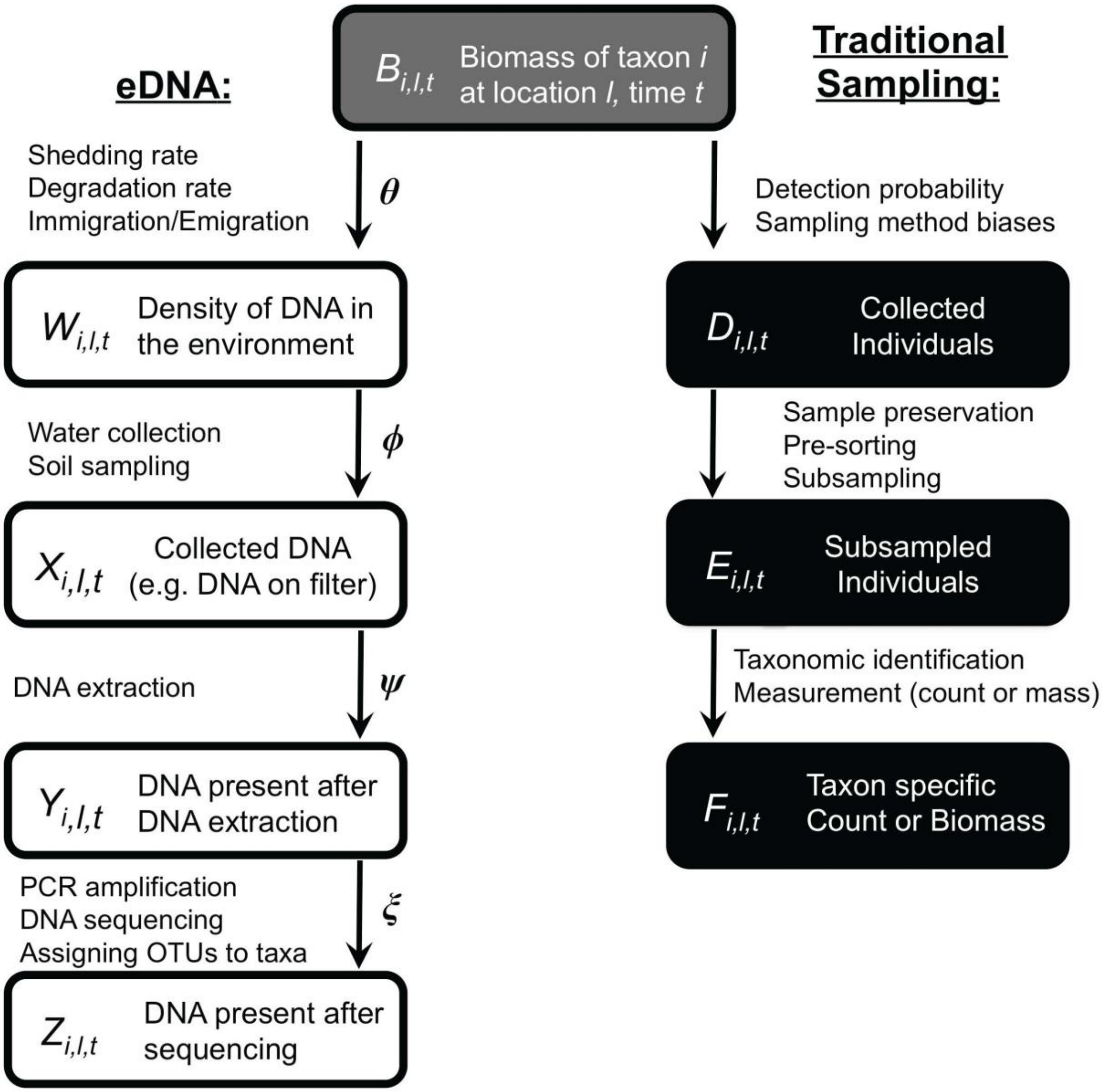
Figure 4: Estimated proportional contribution for 9 unique OTUs and "Other" for three sampled time periods: a) 0930, b) 1130, c) 1230 on June 26, 2014. Points represent raw counts of DNA sequences for each OTU divided by the total number of DNA reads from individual sequencing runs. Each OTU has six observations, three PCR replicates from each of two PCR primers (see text for details). Boxplots show the posterior median, interquartile range and 95% credible interval for each OTU.

Figure 5: Time-series for 3 example OTUs. Median estimated proportion and interquartile range is shown against the tidal height at the sample site (dashed line). Note the left y-axis varies among panels.

Accepted Article

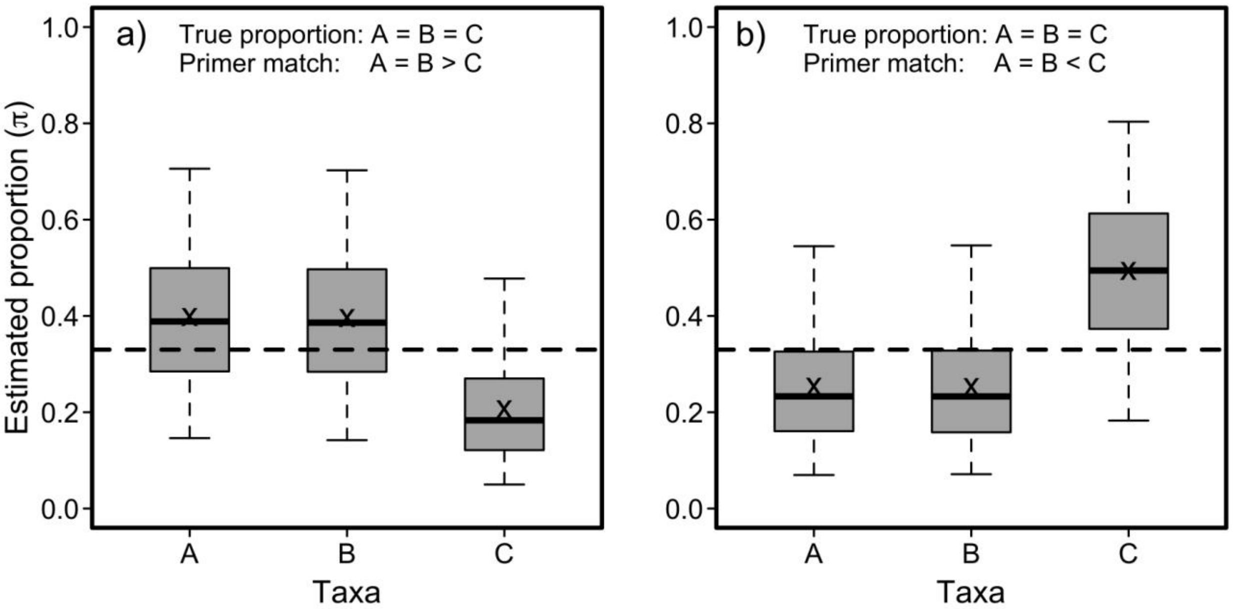
This article is protected by copyright. All rights reserved.

Accepted Article


eDNA:
$B_{i,l,t}$ Biomass of taxon *i* at location *l*, time *t*
Traditional Sampling:
Shedding rate
Degradation rate
Immigration/Emigration
θ
$W_{i,l,t}$ Density of DNA in the environment
Water collection
Soil sampling
ϕ
$X_{i,l,t}$ Collected DNA (e.g. DNA on filter)
DNA extraction
ψ
$Y_{i,l,t}$ DNA present after DNA extraction
PCR amplification
DNA sequencing
Assigning OTUs to taxa
ξ
$Z_{i,l,t}$ DNA present after sequencing
Detection probability
Sampling method biases
$D_{i,l,t}$ Collected Individuals
Sample preservation
Pre-sorting
Subsampling
$E_{i,l,t}$ Subsampled Individuals
Taxonomic identification
Measurement (count or mass)
$F_{i,l,t}$ Taxon specific Count or Biomass

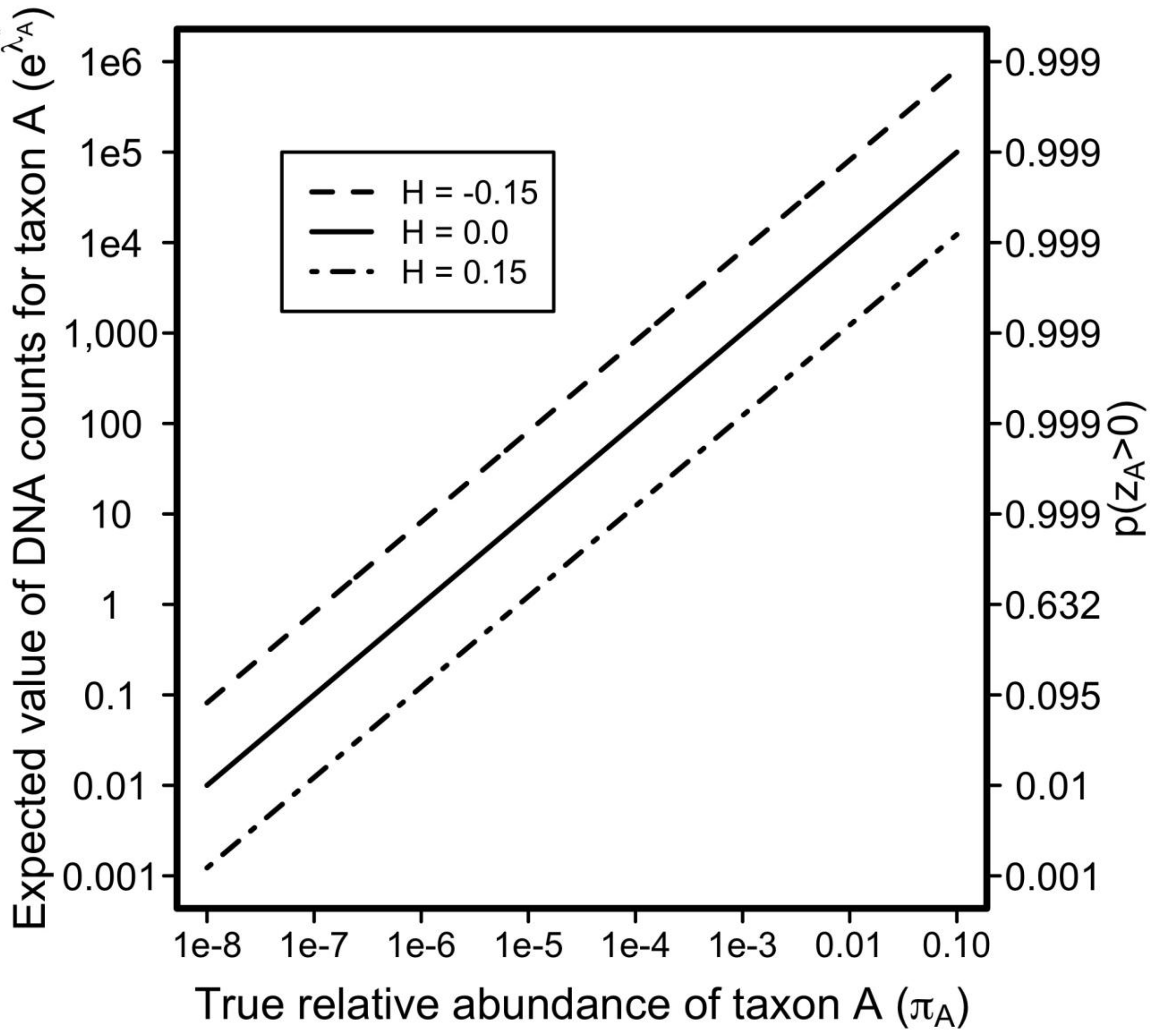
This article is protected by copyright. All rights reserved.

Accepted Article



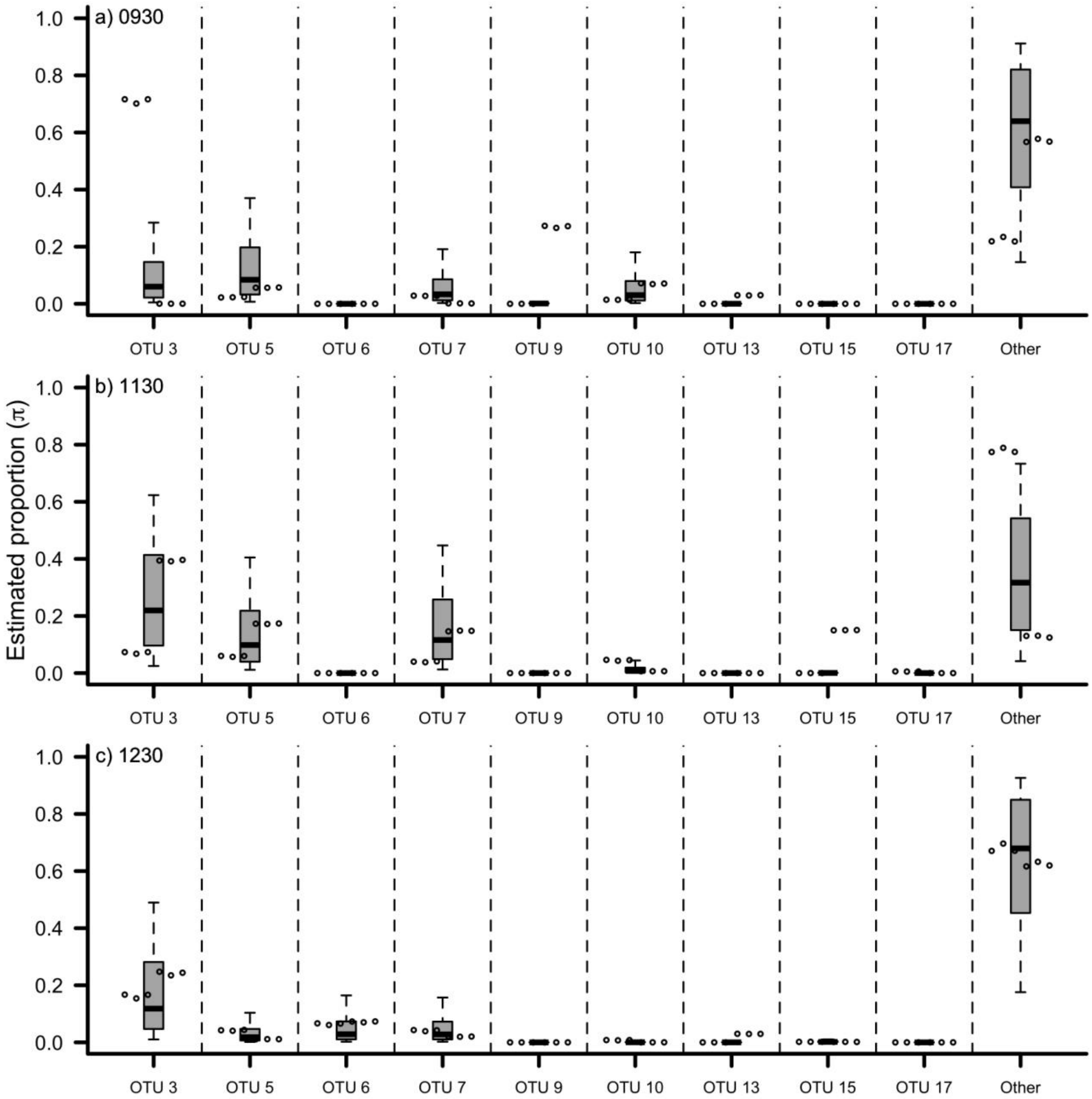
This article is protected by copyright. All rights reserved.

Accepted Article



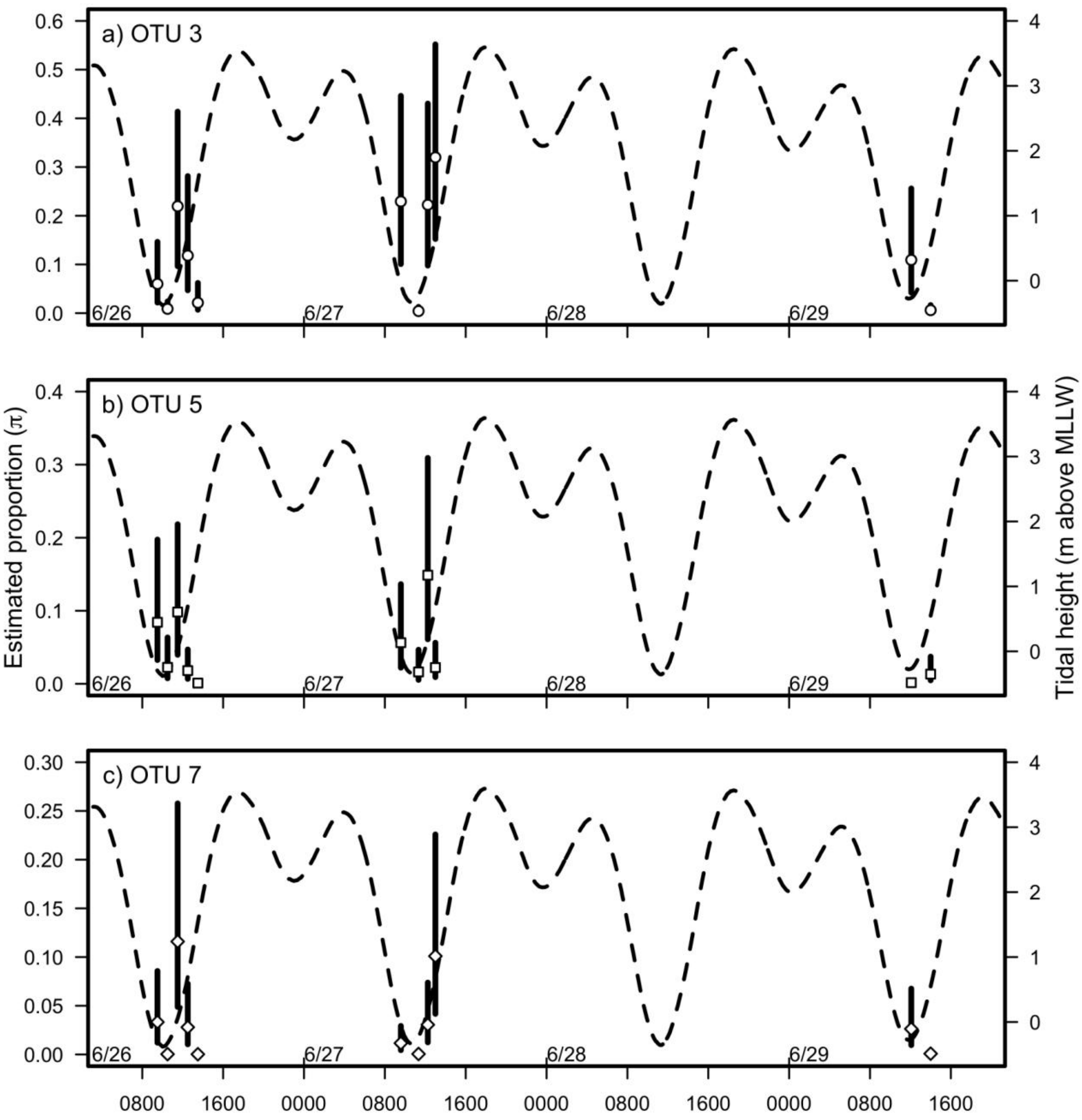
This article is protected by copyright. All rights reserved.

Accepted Article



This article is protected by copyright. All rights reserved.

Accepted Article



This article is protected by copyright. All rights reserved.